Non-homogeneous temporal Boolean models to study endocytosis

M. Ángeles Gallego, M. Victoria Ibáñez and Amelia Simó
Department of Mathematics.
Universitat Jaume I. 12071 Castellón.
mgallego034z@cv.gva.es; mibanez@mat.uji.es; simo@mat.uji.es

July 11, 2011

Abstract

Many medical and biological problems require the analysis of large sequences of microscope images, these images capture phenomena of interest and it is essential to characterize their spatial and temporal properties. The purpose of this paper is to show a new statistical methodology for estimating these parameters of interest in image sequences obtained in the observation of endocytosis. Endocytosis is a process by which cells traffic molecules from the extracellular space into different intracellular compartments. These images are obtained using a very specialized microscopy technique called Total Internal Reflecting (TIRFM).

The Homogeneous Temporal Boolean Model (HTBM) has been recently used to analyze these type of sequences of images. By using a HTBM, spatial homogeneity of events in the cell membrane must be assumed but this is an open question in the biological understanding of the endocytic process. Our aim in this paper is to generalize this methodology to overcome this drawback. In the methodological aspect this work has a threefold aim: to broaden the notion of HTBM by introducing the concept of Non-Homogeneous Temporal Boolean Model; to introduce a hypothesis testing procedure to check the spatial homogeneity assumption; and finally, to reformulate the

existing methodology to work with underlying non-homogeneous point processes. We check the goodness of our methodology on a simulated data set and compare our results with those provided by visual inspection and by assuming spatial homogeneity. The accuracy of the results obtained with simulated data ensure the validity of our methodology. Finally we apply it, as an illustration, to three sequences of a particular type of endocytosis images. The spatial homogeneity test confirms that spatial homogeneity cannot be assumed. As a result, our methodology provides more accurate estimations for the duration of the events and, information about areas of the membrane with higher accumulation of them.

Keywords: Temporal Boolean model, Endocytosis, Spatial non-homogeneity, germ-grain model, parameter estimation.

1 Introduction

There are many practical situations in a wide variety of technological and scientific fields, in which researchers need to manage image data in order to achieve conclusions about a phenomenon of interest. These images are often binary images showing the area covered by a given phenomenon in a certain region. Our paper is concerned with the analysis of endocytosis, a particularly interesting process in cell biology. Endocytosis is a cellular process whereby some materials (e.g. nutrients) are drawn into the cell by means of invagination of the plasma membrane. This process happens in discrete events in which cargo-loaded vesicles detach from the plasma membrane and are trafficked into the cell.

Endocytosis is required for a vast number of vital functions for the well-being of a cell. It regulates many processes, including nutrient uptake, neurotransmission, antigen presentation, pathogen entry, cell adhesion and migration, mitosis, growth and differentiation, and drug delivery.

Although there exists different types of endocytosis, we focus into receptor-mediated

endocytosis that it is a type of endocytosis highly selective because it only includes those molecules (ligands) that bind to the receptor. Within this kind of endocytosis the major endocytic pathway in most cells, and also the best understood, is mediated by the protein clathrin. This protein assists in forming a coated pit on the inner surface of the plasma membrane of the cell. This pit then buds into the cell to form a free clathrincoated vesicle (CCV) in the cell cytoplasm. Coated pits can concentrate large extracellular molecules that have different receptors responsible for the receptor-mediated endocytosis of ligands, e.g. low density lipoprotein, transferrin, growth factors, antibodies and many others.

The life cycle of a CCV involves a sequence of regulated events: a) Cargo loading, where cargo molecules bind to receptors on the plasma membrane; b) Coat assembly, where a molecular lattice of clathrin molecules covers a portion of the plasma membrane containing the cargo-receptor complex; c) Vesicle budding, followed by its pinching-off from the plasma membrane; d) Internalization and coat disassembly; e) Intracellular trafficking of the endocytosed vesicle.

Numerous efforts have recently been made to develop microscopic techniques that allow real-time imaging for endocytosis with a high degree of accuracy. One of these techniques is known as Total Internal Reflection Fluorescence Microscopy (TIRFM) [1, 25]. This technique illuminates a very thin section near the cell-coverslip interface and gives a very high signal-to-background ratio, thus facilitating the visualization of cellular processes near the plasma membrane. Using TIRFM, the assembly of fluorescently labelled clathrin where endocytosis is taking place, results in the appearance of a diffraction-limited spot. The areas of fluorescence generated by different endocytic spots overlap and form random clumps which have different size, shape and duration.

The time which elapses between the appearance and the disappearance of a fluorescent clathrin spot is defined as the duration, or lifetime, of a discrete endocytic event [7].

In cultured cells, the lifetimes of an endocytic event takes aprox. 1 min [8], although

it depends on the type of cell and the size of the cargo particle. It can range from a few seconds to two minutes, even in some exceptional cases, it can be longer than 15 minutes. With respect to the expected frequency of appearance, it also depends on the type of the cell. As an example, it is expected that a 25% of the plasma membrane of a fibroblast is made up of coated pits. With BSC1 cells (kidney epitelial cells of monkey origin), it is expected to find even more than a 95% [10, 7].

Fig. 1 (a) shows several subimages of an endocytic event (highlighted with an arrow mark) which appears (birth) at time 4 s and disappears (death) at time 48 s. Fig. 1(b) plots the brightness profile as a function of time of this endocytic spot.

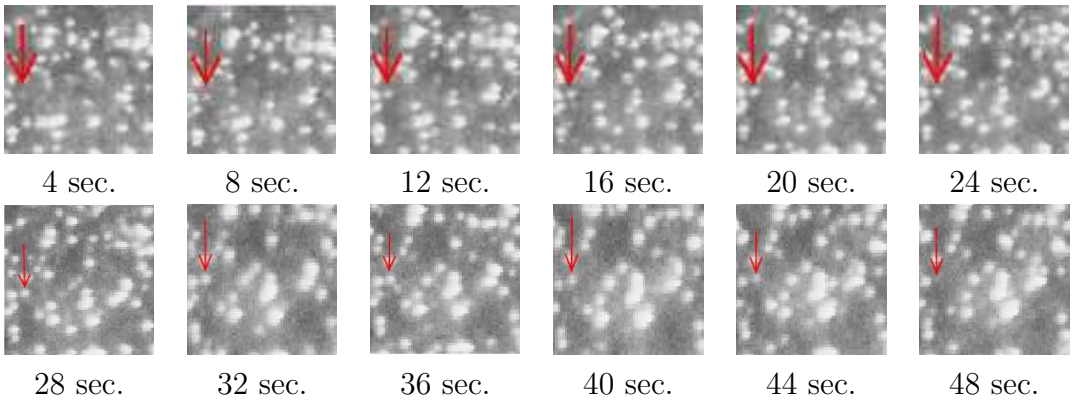
Fig. 5 (b) displays the segmented endocytic spots after image processing of one frame of these sequences. Each connected component may involve one or an unknown number of overlapping CCV.

The spatial and temporal distribution of these clumps is influenced by many biological factors and there is no precise biological knowledge about their spatial distribution in the plasma membrane. In fact, this is one of the questions that remain unsolved in the biological understanding of the endocytic process. A visual inspection of the images is enough to realize that there are certain parts of the plasma membrane with a higher intensity of events.

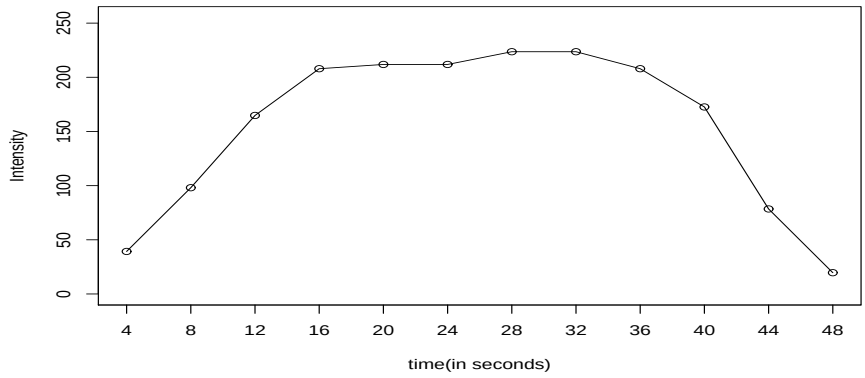
Therefore, to characterize endocytic events it is crucial to estimate certain quantities of interest such as the mean number of endocytic events per unit area and per unit time at different spatial sites and their lifetime [4]. Due to endocytic spots overlapping and clump formation, it is not possible to carry out these tasks in a trivial way.

Although imaging techniques have been widely developed, the obtained image processing is still quite poor. So it is in this context where we focus our work and where the basis of our contribution is.

Due to the importance of clathrin-dependent endocytosis, there has been a great deal of interest among researchers and it has become a very active research field in biological literature over the last decade. Detailed biological models for the production



(a)



(b)

Figure 1: (a) Several subimages of an endocytic event over time. (b) Its brightness profile.

and internalization of clathrin-coated vesicles have been suggested [20, 21, 9]. However, several questions regarding the events and the interactions involved in the endocytic process remain unanswered.

The majority of these studies have been based on a mere visual inspection, or on limited statistical analyses which are typically executed manually, counting one-by-one only certain events: the ones that can be completely observed from the beginning to the end of their life spans and which do not overlap with their neighbours. In other studies these isolated spots are labelled manually and processed using popular image analysis software like MetaMorph and Photoshop (Adobe Systems, San Jose, CA). These methodologies were used in [7, 18, 19] amongst many others.

These "manual" procedures have several limitations. First, each experiment involves dozens of sequences with thousands of frames each, so it is unfeasible to work with large image sequences because there is too much information to process. Second, manual marking of isolated events is quite subjective. Third, the selection of only isolated events leads to a biased sample, i.e. smaller and shorter events are more likely to be included in the sample. In [6] we can find a simulation study showing that the use of this biased sample can lead to a very high bias in the estimation of the time distribution. The greater overlap among grains is, the more biased the estimation will be.

These drawbacks have recently been overcome by Sebastián et al. [22, 2], who use an approach based on the homogeneous Boolean model. This popular stochastic probabilistic model is used to describe the images formed by random clumps that are found in the observation of clathrin-mediated endocytosis dynamics. The Boolean model explicitly considers and accepts this overlap and provides a good description for many irregular patterns observed in microscopy, material sciences, biology, medicine, chemistry, and geostatistics. The model formalizes (in a mathematical sense) the configuration of independent and randomly placed particles. What is observed is a pattern of overlapping random shapes [14].

Mathematically, a homogeneous Boolean model is a random closed set consisting of a Poisson point process in $\mathbf{R}^2$ (called the space) of intensity λ, producing the locations of the germs, coupled with an independent random shape process (i.e. with a sequence of independent and identically distributed random closed sets, called grains). The connected components made of overlapping shapes are called clumps.
A more in-depth study of this model can also be found in [24, 15, 5] and [23]. All these books offer an overview of the different methods that have been developed to overcome the difficulty of estimating the most important parameters of a Boolean model. These parameters are the intensity of events per unit area and the shape distribution of the typical grain. Several methods implemented in applications of Boolean models for real images can be found in [24, 23, 13] and [12].

Sebastián et al. generalize this concept in [2] introducing the notion of homogeneous temporal Boolean model (HTBM). Mathematically, a HTBM is defined as a Poisson point process in $\mathbf{R}^2 \times \mathbf{R}_+$ (space and time) of intensity λ, producing the locations and the births of the germs, coupled with an independent randomly-shaped process (the grains) and an independent time duration process for grains.

In [2] and [22], Sebastián et al. also proposed methodologies to estimate the parameters of interest of the model, such as the duration of the events and their spatial intensity.

Other more mathematically complex generalizations of the Boolean model, such as the germen-grain model or the Gibbs model [24], could theoretically be applied to model the endocitosys but nowadays, it is almost impossible to work with them in an applied context. Because of their complexity, no efficient parameter estimation techniques have been developed yet.

The approach introduced in [22] has an important drawback in our target application: by using a HTBM, spatial homogeneity is being assumed. That would mean, in the endocytosis study context, that the endocytic spots are supposed to be uniformly distributed in the whole cell membrane. As it was explained previously, there is no

scientific evidence to accept this assumption of homogeneity, in contrast, it is believed to be false, and there is a great biological interest in identifying which areas of the cell membrane present a greater accumulation of events.

Although it frequently fails, the spatial homogeneity hypothesis is commonly assumed in most applications of the Boolean model for real problems because it facilitates estimating the parameters of the model.

The novelties introduced by our work are: firstly, the relaxation of the spatial homogeneity hypothesis by introducing the concept of Non-Homogeneous Temporal Boolean Model (NHTBM) as a generalization of [22]; secondly, the introduction of a hypothesis testing procedure that allows us to statistically prove the non-homogeneity hypothesis; and finally, to propose a generalization of the methodology described [22] to efficiently estimate the parameters of interest in the new model.

We apply it to analyze the behavior of the clathrin-dependent endocytic machinery. We reject the hypothesis of spatial homogeneity in the distribution of the events. The use of a model that is more closely adjusted to the physiological characteristics of the real problem leads to more accurate estimators, and it solves one of the open biological questions regarding which parts of the membrane present a greater accumulation of events.

The rest of the paper is organized as follows: the theoretical models and methodologies are introduced in Section 2. In Section 3 a simulation study is carried out to test the performance of the parameter estimation procedure. In Section 4 the methodologies are applied to analyze the dynamics of the GFP-clathrin protein. Finally, in Section 5, conclusions are stated.

2 Models and methods

2.1 Non-homogeneous temporal Boolean model

In this section, we introduce the definition of Non-homogeneous temporal Boolean model, based on the definition of HTBM [2], and the notation that will be used in the rest of the paper.

Definition 1 (Non-homogeneous temporal Boolean model) *Let $\Psi = \{(x_i, t_i)\}_{i\geq 1}$ be a Poisson point process in $\mathbf{R}^2 \times \mathbf{R}_+$, homogeneous in time but non-homogeneous in space, with intensity function $\Lambda(x)$, $x \in \mathbf{R}^2$. Let $\{A_i\}_{i\geq 1}$ be a sequence of independent and identically distributed random compact sets in $\mathbf{R}^2$, and let $\{d_i\}_{i\geq 1}$ be a sequence of independent and identically distributed (as D) positive random variables and that $E\nu_3(A_0 \times [0, D] \bigoplus \check{K}) < +\infty$ for any compact subset K of $\mathbf{R}^3$. Then, the non-homogeneous temporal Boolean model is defined as:*

$$\Phi = \bigcup_{i\geq 1} (A_i + x_i) \times [t_i, t_i + d_i],$$

where $E\nu_3(A_0 \times [0, D] \bigoplus \check{K}) < +\infty$ is a technical condition necessary for the definition that fulfills most real applications. In this formula, E denotes the expectation; for any sets A and B in $\mathbf{R}^3$ $\nu_3(A)$ denotes the volume of A, and $A \bigoplus B$ denotes their Minkowsky addition.

The great difference between the definition of HTBM and NHTBM lies in the fact that the intensity of the events given by the constant parameter λ in the homogeneous case, now becomes a function $\Lambda(x)$ of the spatial sites $x = (x^{(1)}, x^{(2)})$ (the intensity function).

Figure 2 shows several frames corresponding to simulations of this model with intensity function $\Lambda(x) = \lambda_0 x^{(1)} x^{(2)}$, assuming circular grains with random uniform radii and with two types of random duration: uniform and exponential. Different

images correspond with different values for λ_0 and for the parameters of the radii and duration distributions. These sequences of images will be used in the simulation study implemented in Section 3.

In the applications, we will work with binary images sequences that will be considered as realizations of an NHTBM. So, they will be considered as samples of a spatiotemporal infinite process, as defined in Def. 1. The spatiotemporal sampling window will be denoted by $W \times [0, T]$ and the sampling times will be denoted by $s_1 < s_2 < \cdots < s_m$, with $0 \leq s_1$; $s_m \leq T$. Then, the observed data set will be:

$$\{\Phi_{s_i}\}_{i=1,\cdots,m} \quad \text{with} \quad \Phi_{s_i} = \Phi \bigcap (W \times \{s_i\}) \quad \forall i = 1, \cdots, m \tag{1}$$

i.e. a discrete set of temporal cross-sections of the model, corresponding to the observation times s_i, $i = 1, \cdots, m$.

2.2 Parameter estimation

In this section we propose the methodology required to estimate the parameters of interest of a NHTBM. Our aim is twofold: to estimate the intensity function of the germ process, $\Lambda(x)$, and to estimate the probability distribution of the durations of the events i.e. the probability distribution of random variable D.

Two different approaches are found in [2] and [22] to manage parameter estimation in a HTBM. Both of them are based on the analysis of the variation in the intensity of the germ process throughout time but only the second approach can be generalized to the non-homogeneous case. This approach uses several cross-section aggregations to analyze the increase in intensity, and from now on, we are going to follow it, so the following sequences are defined:

$$\widetilde{\Phi}_{s_i} = \bigcup_{j=i}^{i+k} \Phi_{s_j} \quad i = 1, \cdots, m-k. \tag{2}$$

In these sequences, the grains size will keep its original distribution, although the spatial intensity for the germs process will be higher. This rise in intensity for the aggregated model will only depend on the number of frames aggregated and their time lags (time delay between two aggregated frames). The parameters of interest of our model will be estimated by analyzing these increases in intensity.

Some previous results are needed in order to obtain the desired estimates.

The first of these previous results is related with the temporal cross-sections of the model and is a generalization of the proposition 1 of Ayala et al [2].

Proposition 1 *If Φ is an NHTBM with intensity function $\Lambda(x)$ and primary grain $A_0 \times [0, D]$, then the temporal cross-section $\Phi_s = \Phi \cap (\mathbf{R}^2 \times \{s\})$ is a non-homogeneous Boolean model in $\mathbf{R}^2$ with primary grain A_0 and intensity*

$$\Lambda_s(x) = \Lambda(x)ED. \tag{3}$$

The proof of proposition 1 is trivial following similar arguments to those given in [2].

The second theoretical result is related to the aggregated frames and it is also a generalization of those given in [22].

Proposition 2 *If Φ is an NHTBM with intensity $\Lambda(x)$, each $\widetilde{\Phi}_{s_i}$ defined as in eq. (2) is a realization of a non-homogeneous Boolean model in $\mathbf{R}^2$, and if the times s_i are equally spaced, with $s_i - s_{i-1} = \delta$, $\forall i = 2, \cdots, m$, the intensity function of $\widetilde{\Phi}_{s_i}$, $\lambda_s(k, \delta, x)$, is*

$$\lambda_s(k, \delta, x) = \Lambda(x)\Big[kp(0) - (k-1)p(\delta)\Big] \tag{4}$$

where

$$p(s) = \int_s^{+\infty} P(D \geq t)dt; \quad p(0) = \int_0^{+\infty} P(D \geq t)dt = ED \tag{5}$$

Note that the intensity of aggregated frames only depends on the time between consecutive frames δ, and the number of accumulated frames k. The proof of proposition 2 can be found in the appendix.

After introducing these theoretical results we are in a position to propose a methodology for estimating the parameters of the NHTBM.

The first step in the procedure consists in the estimation of $\lambda_s(k, \delta, x)$, the intensity function for each $\widetilde{\Phi}_{s_i}$. Proposition 1 tell us that each $\widetilde{\Phi}_{s_i}$ is a realization of a spatial non-homogeneous Boolean model. The estimation of the intensity function is much more complex for the non-homogeneous case than for the homogeneous one and algorithms in the non-homogeneous literature are scarce. Among the existing procedures, we propose the estimation procedure given by Molchanov and Chiu [16], which is based on the estimation of the intensity of tangent points process using kernel techniques combined with kernel estimation of the coverage function. This procedure will be repeated for different values of k and δ.

Once $\lambda_s(k, \delta, x)$ has been estimated we follow the ideas stated in [22].

Equation 4 tells us that for each x and for each δ, $\lambda_s(k, \delta, x)$ is a linear function on k, if we denote $\alpha(\delta, x)$ and $\beta(\delta, x)$ for the slope and constant parameter of this linear function we have:

$$\alpha(\delta_0, x) = \Lambda(x)\Big[p(0) - p(\delta_0)\Big],$$
$$\beta(\delta_0, x) = \Lambda(x)p(\delta_0).$$

And for each value of x:

$$\alpha'(0, x) = -\Lambda(x)p'(0) = \Lambda(x)P(D \geq 0) = \Lambda(x) \tag{6}$$

$$\alpha''(\delta, x) = \Lambda(x)f_D(\delta) \tag{7}$$

Fitting a linear function to the estimates $\hat{\lambda}_s(k, \delta, x)$, numerically deriving with re-

spect to δ and substituting in eq. 6 we obtain the desired estimate of the intensity function $\hat{\Lambda}(x)$.

Analogously, taking into account eq. 7, we obtain an estimate of the probability density of D for each site:

$$\hat{f}_D(\delta) = -\frac{1}{\hat{\Lambda}(x)}\hat{\alpha}''(\delta, x) \tag{8}$$

We use their mean as the final estimate of the probability density function.

Finally combining eq. 3 with $\hat{\Lambda}(x) = \hat{\alpha}'(0, x)$ we get an estimate of ED:

$$\hat{ED} = \frac{1}{\sharp W}\sum_{x \in W}\left[\frac{\frac{1}{m}\sum_{j=1}^{m}\hat{\Lambda}_{s_j}(x)}{\hat{\Lambda}(x)}\right] \tag{9}$$

2.3 A simple test for spatial homogeneity

In this section we introduce a very simple statistical hypothesis testing procedure that allows us to test the null hypothesis of spatial homogeneity against the alternative of non-homogeneity.

There are no formal homogeneity tests in spatial random sets and point processes literature. Usually, a single observation in a sample window is available for the analysis and homogeneous spatial patterns could look like non-homogeneous depending on the size of the window.

Nevertheless, advantage can be taken in temporal models because of the fact that there are several, yet dependent, realizations of the spatial pattern. These dependent realizations can provide us with extra information that is not usually available in this context and which will allow us to perform the statistical hypothesis testing procedure as follows:

- Use the Molchanov method explained in the previous section [16] to estimate the intensity function from each frame.

- Use a batch-means type method [11] to obtain independent replications based on the temporal observations.
- Under the null hypothesis of homogeneity, the estimated value does not depend on the coordinate. Apply the Friedman non-parametric ANOVA test to the sample obtained in the previous step, to compare the estimated values at each position of each frame.

The adequacy of this procedure will be tested in the following section.

3 Simulation study

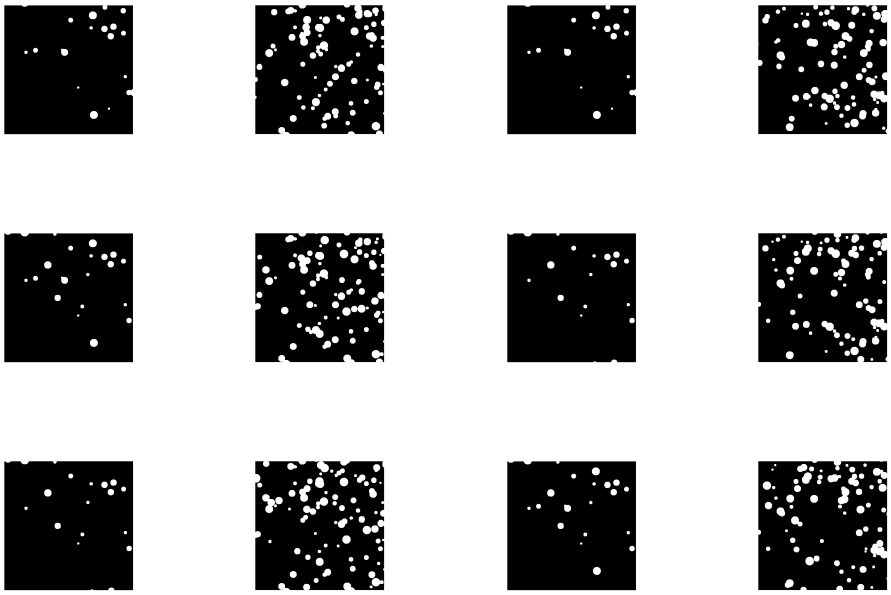
In this section we will set out a simulation study in order to test the performance of the parameter estimation and the homogeneity testing procedures.

It is not possible to evaluate the performance of our methodology without assuming a stochastic model of the (random) generating mechanism for the locations of germs, times of occurrence (birth times) and durations of the events.

We simulate a NHTBM in a 512×512 window, assuming for the primary grain, random discs with uniform radii and an intensity function $\lambda(x) = \lambda_0 x_1 x_2$, so that the two coordinates of germ points are independent. Simulations from two different values for λ_0 ($\lambda_0 = 1.0e-07$ and $\lambda_0 = 5.0e-07$) are obtained. This intensity function was also used by Molchanov in [16]. For each intensity value, two different probability distributions, uniform and exponential, were used for the random duration, both of them with a mean of 15 seconds.

Ten videos of 150 frames each were generated for the different experimental setups described above. The sampling ratio was one frame per second. Figure 2 shows three frames of a typical realization of each model.

As stated above, prior to estimating the parameters of the Boolean model, we checked the performance of the homogeneity test proposed in Section 2.3. To check it,



(a) (b) (c) (d)

Figure 2: Each column shows three consecutive frames of a simulation of a NHTBM in a 512×512 window. First and second column with exponential random duration and (a) $\lambda_0 = 1.0e-07$ and (b) $\lambda_0 = 5.0e-07$. Third and fourth column with uniform random duration and (c) $\lambda_0 = 1.0e-07$ and (d) $\lambda_0 = 5.0e-07$.

we applied this test to all the sequences of simulated images. Different sample sizes were used to obtain batch means samples and, in all cases and all simulated sequences, the p-value obtained from the Friedman test was almost equal to 0. Thus, the hypothesis of homogeneity was clearly rejected in all cases.

The Molchanov method was used to estimate the intensity function of each 2D realization and, for smoothing purposes, we used the Epanechnikov kernel with bandwidths of $h = 84$ for $\lambda_0 = 1.0e-07$ and $h = 80$ for $\lambda_0 = 5.0e-07$. The same value of h was used for both the density of the exposed tangent point process and the coverage function. These choices of the bandwidth correspond to the means of the optimal for 10 realizations. We used an approach based on functional data analysis [17] in order to obtain a more precise and smoother estimation of $f_D(\delta)$. Functional data analysis converts the raw data $\hat{\alpha}(\delta, x)$ into a functional form for each x, and so both it and its derivatives can be evaluated in a more precise way at all values over an interval in time. A basis must be specified to achieve our aim and among several possibilities, we chose a polynomial spline basis with 8 basis functions of order 5 (implying piecewise fourth-order polynomials). Order 5 was used to obtain a smooth second-order derivative.

We wrote a library of functions to be used to carry out the different tasks in MATLAB[1].

The estimation of the spatial intensity function in all cases, can be seen in Fig 3. This figure compares the theoretical density functions for the two values of λ_0 with the estimates obatined for the 4 simulated sequences of images. New images that show the differences between theoretical and estimated functions are also included in Fig 3. Regarding the estimation of the density function of the duration of the events, Figures 4 (a) and (b) show the theoretical density functions for the different simulation scenarios, and the mean of the estimates obtained for these temporal density functions.

In order to give a numeric measure of the errors in the estimation of the intensity

[1]MATLAB is a trademark of The MathWorks Inc.

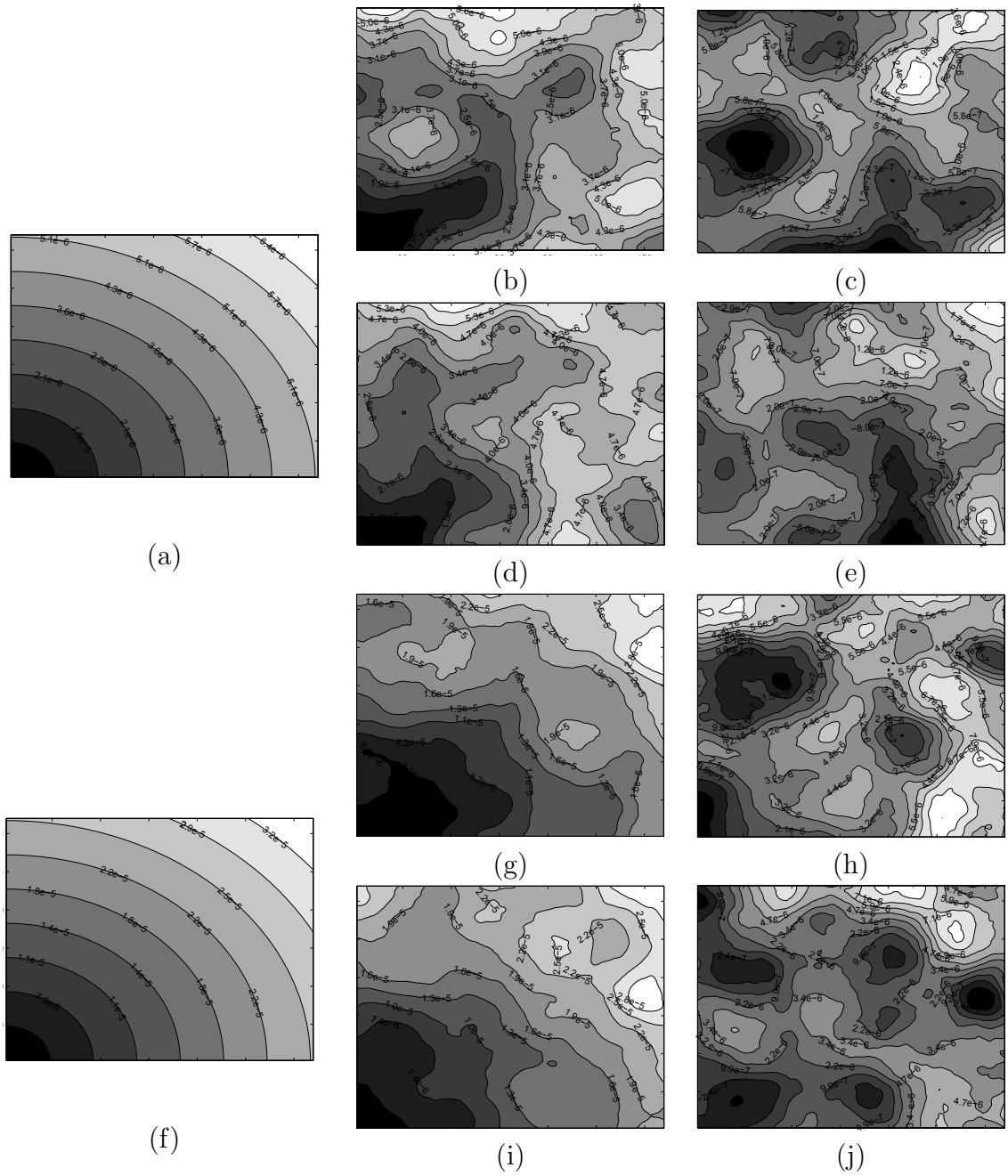


Figure 3: Results of the simulation to estimate the intensity functions. In all the images, the darker the color, the lower the intensity is. First column: theoretical intensity functions with (a) $\lambda_0 = 1.0e-07$ and (f) $\lambda_0 = 5.0e-07$. Second column: estimation of the intensity functions . (b) Estimation of the sequence simulated with $\lambda_0 = 1.0e-07$ for the exponential distribution, (d) with $\lambda_0 = 1.0e-07$ for the uniform distribution, (g) with $\lambda_0 = 5.0e-07$ for the exponential distribution and (i) with $\lambda_0 = 5.0e-07$ for the uniform distribution. Third column (figs (c),(e),(h) and (j)): contour plots of the differences between first and second column (theoretical and simulated intensity functions).

function, we calculated the mean square errors of the estimations over all the positions and simulations. To get a relative measurement, we obtain their square root and divide them by their mean. Results are given in Table 1. In the two first rows of Table 2 the average and standard deviation of the estimates of the mean of durations are given. The results obtained with our method are completely satisfactory. As can be seen, better values are obtained for lower intensity values. This can be due to the own features of the tangent point method.

	$\lambda_0 = 1.0e-07$	$\lambda_0 = 5.0e-07$
Exponential	0.2575	0.2301
Uniform	0.2187	0.1795

Table 1: Relative errors in the estimation of the Intensity function for the different temporal distributions and λ_0 values.

Once our methodology has been proposed and explained, we would like to compare the results that we have obtained with the ones obtained analyzing the same image sequences with other methods existing in the literature, which deal with the same kind of information.

As stated above, the main novelty of our methodology is its ability to estimate the spatial intensity of events per unit area and time at different spatial sites and also their life time. Other methods existing in the literature, such as working with manually labeled isolated points or the methodology proposed by Sebastián et al. [22, 2], assume spatial homogeneity and therefore, they provide a single value as an estimation of the "common" spatial intensity. This makes it impossible to compare the estimate of the spatial intensity obtained with our method (a 2D function of the spatial site as shown in Fig 3) to the estimate obtained with the other methods (a single number). On the other hand, all these methodologies (manual labeling, HTBM and NHTBM) assume temporal homogeneity, so all the estimates of the mean life time of the events obtained with the three methodologies can be compared.

Only simulated images with $\lambda_0 = 1.0e-07$ will be used for the comparison. The

lower the spatial intensity is, the easier it will be for us to identify, label and analyze isolated events. On the other hand, as Ayala et al. [6] say, the use of this biased sample can lead to a very high bias in the estimation of the distribution of the duration. The greater overlap among grains is, the more biased the estimation will be.

Table 2 also shows the mean and standard deviation of the estimates for the mean of the durations of the events, obtained with the three methods, for both the uniform and exponential temporal distributions. Figure 4 also shows the temporal density function for both distributions. Figures 4(c) and 4(d) show the theoretical density functions (solid lines), the density functions estimated by using our method (dotted lines), those estimated from manually identified and processed isolated points (dashed lines), and those estimated by using the method proposed by Sebastián et al. (dash-dot lines). As it can be seen, the results obtained with our method do not differ too much of those obtained with the methods nowadays in use. Regarding the estimation of the mean of the duration of the events (Table 2), they represent a slight improvement. In the estimates obtained for the uniform distribution, our method obtains the closest average of the estimates of ED to the real value. We also obtained a value for the standard deviation much smaller than the obtained with the rest of methods. In the case of the exponential distribution, although the mean estimated with our method differs a bit more of the real value than the obtained with the other methods, the standard deviation continues being smaller, which means that the estimations obtained with our method are more reliable than the obtained in the other cases.

4 Application

In this section we show the application of our models and methods to sequences of images used to analyze clathrin-mediated endocytosis dynamics.

As an illustration, we are going to work with image sequences derived from three videos of COS-7 monkey fibroblast cells. These cells themselves most resemble fibrob-

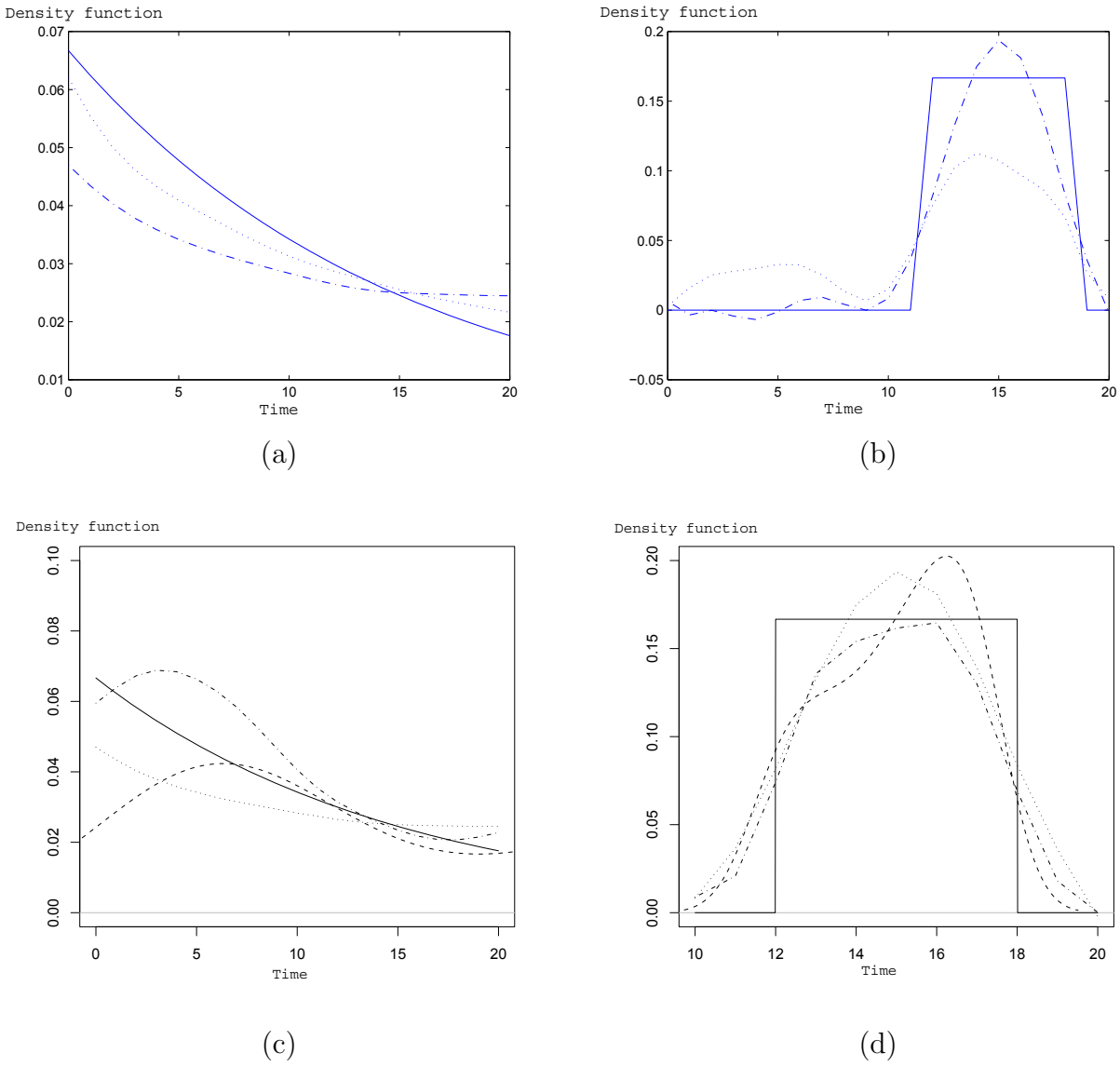


Figure 4: (a) Temporal density function for exponential duration. Theoretical value (solid line); estimated function for $\lambda_0 = 1.0e-07$ (dash-dot line) and estimated function for $\lambda_0 = 5.0e-07$ (dotted line). (b) Temporal density function for uniform duration. Theoretical value (solid line); estimated function for $\lambda_0 = 1.0e-07$ (dash-dot line) and estimated function for $\lambda_0 = 5.0e-07$ (dotted line). (c) Temporal density function for exponential duration. Theoretical density function (solid line); estimate density function by using our method (dotted line); from manually identified and processed isolated points (dashed line) and by using the method proposed by Sebastián et al. (dash-dot line). (d) Temporal density function for uniform duration. Theoretical density function (solid line); estimate density function by using our method (dotted line); from manually identified and processed isolated points (dashed line) and by using the method proposed by Sebastián et al. (dash-dot line).

			mean ED	standard dev. ED
$\lambda_0 = 5.0e-07$	Our method	Uniform	16.47	0.26
		Exponential	18.38	0.30
$\lambda_0 = 1.0e-07$	Our method	Uniform	14.92	0.28
		Exponential	15.99	0.73
$\lambda_0 = 1.0e-07$	Isolated points	Uniform	14.09	0.38
		Exponential	15.48	2.34
$\lambda_0 = 1.0e-07$	Sebastian et al.	Uniform	12.56	4.7
		Exponential	14.16	5.64

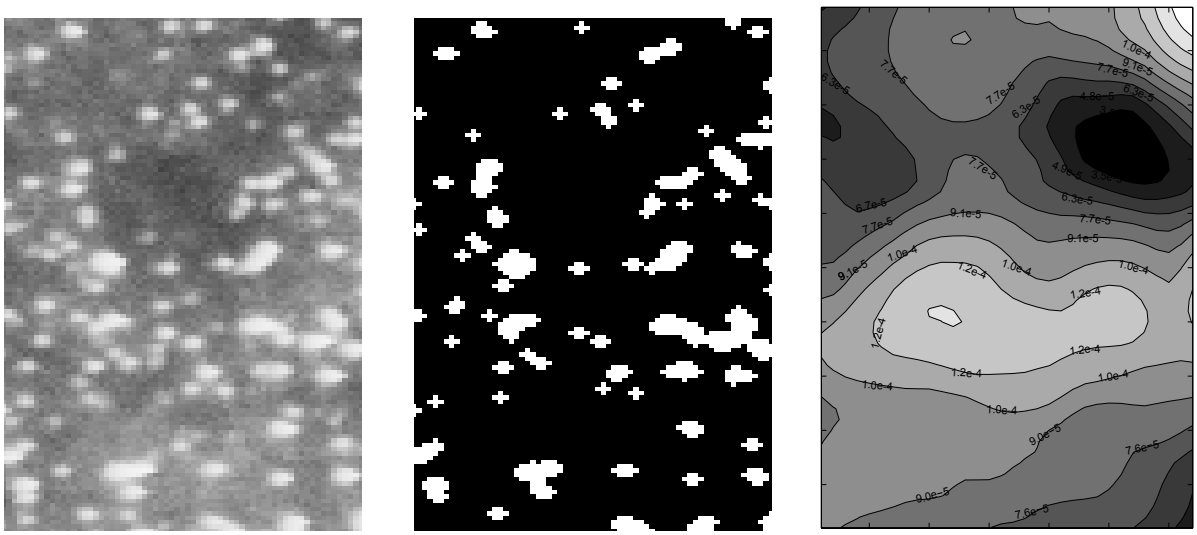
Table 2: Mean and standard deviation of the estimates of the expectation of durations for the different temporal distributions and lambda values. Two first rows show our results. Third and fourth rows show the results obtained using isolated points and Sebastian et al. methods respectively

last cells in humans that are the most common cells of connective tissue. COS is a cell line often used by biologists in cell biology experiments. Cells expressed clathrin light chain coupled to the green fluorescent protein (GFP). Each one of these sequences consists of 300 frames acquired at one frame every four seconds. These sequences correspond to sequences 1, 3 and 5 used in [22].

Other kind of sequences of endocytic images could be analyzed using our methodology provided that the homogeneity test of section 2.3 confirms that does not fulfil the homogeneity assumption.

As a previous step, several pre-processing algorithms must be applied to obtain noise-free segmented images. A detailed explanation of this procedure can be found in [22]. After applying the pre-processing algorithms, our data set consists of three sequences of binary images displaying the segmented endocytic spots. Figure 5(a) shows an image of one of the original sequences, while Figure 5(b) shows the same image once processed and segmented.

Once the images have been segmented, the first necessary step in our methodology consists in checking whether our data can be really considered as a realization of a non-homogeneous process, therefore, we apply the homogeneity test proposed in Section



(a) (b) (c)

Figure 5: (a) A frame of a sequence of 300 TIRFM images of a cell expressing fluorescent clathrin protein; (b) segmented endocytic spots after imaging processing of the frame showed in (a); (c) Estimated spatial intensity function for the Cell 2, obtained from the full sequence.

2.3. With respect to the estimation of the intensity function of each 2D realization we use again the Epanechnikov kernel. The optimal bandwidth for the kernel was chosen by using the normal approximation as described in [3]. The edge effects near the boundary of the image were corrected by simply reducing the window. On the other hand the sample size for the batch means method was chosen as equal to 5. The results of applying the Friedmand test for each sequence are shown in Table 3, we can find the Chi-square statistic and their corresponding p-values to contrast the null hypothesis of homogeneity. P-values are always almost zero and so the homogeneity hypothesis is clearly rejected. Therefore we can assume our data set as a realization of a NHTBM.

As said above, the use of NHTBM can provide a powerful tool to analyze the behavior of the clathrin-dependent endocytic machinery. On the one hand using NHTBM allows us to estimate the intensity function, so that information about the spatial distribution of the events in the whole membrane is obtained. This fact has a great

Cell	Chi-sq	p-value
1	17137.41	0
2	34438.24	0
3	38600.42	0

Table 3: Results of applying the Friedman test following the methodology explained in Section 2.3.

interest from the biological point of view. Obtaining that information has not been possible up to now using the existing methods. On the other hand, NHTBM allows us to estimate the probability distribution for the duration of events.

Fig. 5(c) shows the spatial intensity function estimated for one of the analyzed cells (Cell 2). This estimation has been obtained from the full sequence of images available for this cell. We can clearly observe a greater density of endocytic spots in the image centre.

As regards to distribution of event durations, Table 4 shows the estimation of mean duration of endocytic events for the three analyzed cells, while Fig. 6 shows the estimation of the density function of event durations for Cell 2. Comparing our results to the ones obtained by Sebastin et al [22], it can be seen that our procedure provides lower values when estimating the mean duration of the endocytic events.

Cell	1	2	3
$\hat{ED}$	37.95	33.39	40.17

Table 4: Estimates of the mean of the durations.

Due to the fact of working with real images, we can not compare our results with the true values of the parameters, since these are unknown. However, it is expected that our results improve theirs for two reasons. On the one hand, because our methodology is based on a model that fits better to the features of the studied process, taking into account the spatial non-homogeneity of our images (that has been tested). On the other hand, from the study with simulated data in Section 3, we found that our

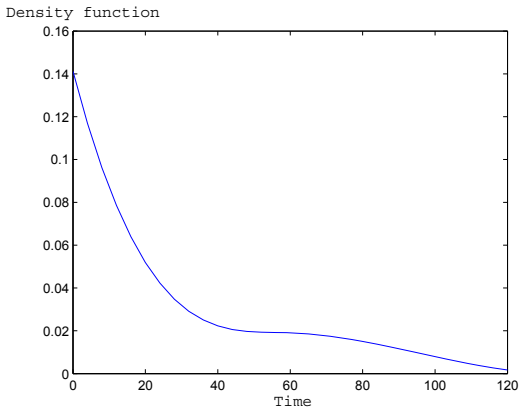


Figure 6: Estimated density functions of durations for Cell 2.

methodology yields better results in a situation of spatial nonhomogeneity.

We have shown a methodological tool that can be applied in the study of endocytosis but a deeper study conducted jointly with experts in biological sciences and a larger sample would be necessary to achieve biological conclusions.

5 Conclusions

In this paper we have proposed both a probabilistic model and a statistical methodology that generalize the methodology proposed by Sebastián et al. [22] to study the kinetics of endocytosis in living cells. The novelty of our approach is the relaxation of spatial homogeneity hypothesis by introducing the concept of a non-homogeneous temporal Boolean model. The homogeneity assumption is a common hypothesis in most applications because it facilitates estimating the parameters of the Boolean model, although it fails when the spatiotemporal distribution of endocytic spots is analyzed. In fact, this is one of the questions that remain unsolved in the biological understanding of the endocytic process.

Using formulas obtained in Section 2.2, to estimate the parameters of the model requires the application of previous statistical techniques and methodologies, such as

functional data analysis [17], to estimate the spatial intensity function of 2D non-homogeneous Boolean models [16]. Moreover, these previous techniques depend on a large number of tuning parameters which often have a great effect on the results. As an example, the adequacy of Molchanov's method to estimate the intensity function [16] depends on the choice of the bandwidth, and may not be very satisfactory when there are few tangent points. Despite everything, the results obtained in the simulation study are quite satisfactory. On the other hand, more precise techniques for working with functional data would also be advantageous in order to obtain better fits. Nonetheless, our methodology opens up a door to achieve the analysis of a great number of real applications where the underlying spatial process clearly does not fulfil the homogeneity assumption, like the sample that has been analyzed in this work.

Regarding the application to endocytosis, we have detected that there are parts of the cellular membrane with a higher accumulation of endocytic spots and we obtain slightly lower estimates for the durations of the endocytic events than the obtained with the methods nowadays in use. A deeper study conducted in conjunction with experts in biology and with a larger data set of image sequences would be necessary to reach biological conclusions.

6 Acknowledgments

We would like to thank Dr. Maria Elena Diaz from the Department of Computer Science of the University of Valencia and Dr. Guillermo Ayala from the Department of Statistics of the University of Valencia for introducing us in this interesting problem, and to Dr. Derek Toomre and Roberto Zoncu from the Department of Cell Biology of Yale University for obtaining the images and allowing us to use them.

This work has been supported by the Spanish Ministry of Science and Education, projects TIN2007-67587 and TIN2009-14392-C02-01, and by the Fundació Caixa Castelló BANCAIXA, project P11A2009-02 .

References

[1] D. Axelrod. Total internal reflection fluorescence microscopy in cell biology. *Traffic*, 2:764–774, 2001.

[2] G. Ayala, R. Sebastián, M.E. Díaz, E. Díaz, R. Zoncu, and D. Toomre. Analysis of spatially and temporally overlapping events with application to image sequences. *IEEE Transactions on Pattern Analysis and machine intelligence*, 28(10):1707–1712, 2006.

[3] A.W. Bowman and A. Azzalini A. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations.* Oxford University Press, 1997.

[4] S.D. Corner and S.L. Schmid. Regulated portals of entry into the cell. *Nature*, 4:37–44, 2003.

[5] N.A.C. Cressie. *Statistics for spatial data.* Wiley Series in Probability and Mathematical Statistics, 1993. (pages 753-775).

[6] M. E. Díaz, G. Ayala, and E. Díaz. Estimating the duration of overlapping events from image sequences using cylindrical temporal boolean models. *Journal of Mathematical imaging and vision*, 38(2):83–94, 2010.

[7] M. Ehrlich, W. Boll, A. van Oijen, R. Hariharan, K. Chandran, M. Nibert, and T. Kirchhausen. Endocytosis by random initiation and estabilization of clathrin-coated pits. *Cell*, 118:591–605, 2004.

[8] I. Gaidarov, F. Santini, R.A. Warren, and J.H. Keen. Spatial control of coated-pit dynamics in living cells. *Nature Cell Biology*, 1:1–7, 1999.

[9] T. Kirchhausen. Clathrin adaptors really adapt. *Cell*, 109:413–416, 2002.

[10] T. Kirchhausen. Imaging endocytic clathrin structures in living cells. *Trends in Cell Biology*, 19:596–605, 2009.

[11] A.M. Law and W.D. Kelton. *Simulation modelling and analysis.* McGraw Hill, 1991.

[12] T. Lyman. *Metals Handbook*. American Society for Metals, 1972.

[13] R. Margalef. *Ecología*. Omega. Barcelona, 1974.

[14] G. Matheron. *Random Sets and Integral Geometry*. J. Wiley & Sons, New York, 1975. (pages 54-155).

[15] I. Molchanov. *Statistics of the Boolean model for practitioners and mathematicians*. J. Wiley & Sons, New York, 1997.

[16] I.S. Molchanov and S.N. Chiu. Smoothing techniques and estimation methods for nonstationary boolean models with applications to coverage processes. *Biometrika*, 87(2):265–283, 2000.

[17] J.O. Ramsay and B.W. Silverman. Functional data analysis. second edition. In *Springer Series in Statistics*. 1997.

[18] J.Z. Rappoport, K.P. Heyman, S. Kemal, and S.M. Simon. Dynamics of dynamin during clathrin mediated endocytosis in pc12 cells. *PLoS ONE 3(6): e2416. doi:10.1371/journal.pone.0002416*, 2008.

[19] J.Z. Rappoport and S.M. Simon. Real time analysis of clathrin mediated endocytosis during cell migration. *Journal of cell science*, 116:847–855, 2002.

[20] S.L. Schmid. Clathrin-coated vesicle formation and protein sorting: an integrated process. *Annu. Rev. Biochem.*, 66:511–548, 1997.

[21] S.L. Schmid. Clathrin-mediated endocytosis: membrane factors pull the trigger. *Trends Cell Biol.*, 11:385–391, 2001.

[22] R. Sebastián, E. Díaz, G. Ayala, M.E. Díaz, R. Zoncu, and D. Toomre. Studying endocytosis in space and time by means of temporal boolean models. *Pattern Recognition*, 39(11):2775–85, 2006.

[23] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, London, 1982. (pages 481-502).

[24] D. Stoyan, W.S. Kendall, and J. Mecke. *Stochastic Geometry and its applications.* Chichester John Wiley & Sons, Second Edition, 1995. (pages 65-95).

[25] D. Toomre and D.J. Manstein. Lighting up the cell surface with evanescent wave microscopy. *Trends Cell Biol*, 11:298–303, 2001.

A Appendix

In this appendix we show the proof of the proposition 2.

Let us define $\Upsilon = \{t_i\}_{i\geq 1}$ as the marginal temporal Poisson point process of the birth time, with intensity $\int_W \Lambda(x)dx$; and $\Upsilon_{s_i} := \{t_n \in \Upsilon : t_n \leq s_i \leq t_n + d_n\}$. It is trivial to prove that the mean number of points in $\bigcup_{j=i}^{i+k} \Upsilon_{s_j}$ (the union of all points alive at some of the k temporal cross-sections) is:

$$\int_W \Lambda(x)dx\Big[kp(0) - (k-1)p(\delta)\Big],$$

with $p(s)$ and $p(0)$ as in eq. (5).

On the other hand, as the mean number of points in $\bigcup_{j=i}^{i+k} \Upsilon_{s_j}$ is equal to the mean number of points in $\widetilde{\Phi}_{s_i}$, and by definition this is equal to $\int_W \lambda_s(k,\delta,x)dx$, then

$$\int_W \lambda_s(k,\delta,x)dx = \int_W \Lambda(x)dx\Big[kp(0) - (k-1)p(\delta)\Big] \tag{10}$$

and as Equation (10) holds for all $W \subset \mathbf{R}^2$, it can be concluded that

$$\lambda_s(k,\delta,x) = \Lambda(x)\Big[kp(0) - (k-1)p(\delta)\Big] \quad \forall x \tag{11}$$